DOCUMENT RESUME

ED 359 229 TM 019 998

AUTHOR Shavelson, Richard J.; And Others
TITLE Sampling Variability of Performance Assessments. Report on the Status of Generalizability Performance: Generalizability and Transfer of Performance Assessments. Project 2.4: Design Theory and Psychometrics for Complex Performance Assessment in Science.
INSTITUTION National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY California Univ., Berkeley. Office of the President.; National Science Foundation, Washington, D.C.; Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE Jan 93
CONTRACT NSF-SPA-8751511; NSF-TPE-9055443; R117G10027
NOTE 32p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; *Educational Assessment; Error of Measurement; Evaluators; *Generalizability Theory; Interrater Reliability; Job Performance; Mathematics Achievement; Measurement Techniques; *Sampling; Science Instruction; *Scores; Scoring; Student Evaluation; *Test Reliability; Test Validity; Training
IDENTIFIERS *Performance Based Evaluation; Science Achievement

ABSTRACT

In this paper, performance assessments are cast within a sampling framework. A performance assessment score is viewed as a sample of student performance drawn from a complex universe defined by a combination of all possible tasks, occasions, raters, and measurement methods. Using generalizability theory, the authors present evidence bearing on the generalizability (reliability) and convergent validity of performance assessments sampled from a range of measurement facets, measurement methods, and data bases. Results at both the individual and school level indicate that rater-sampling variability is not an issue: raters (e.g. teachers, job incumbents) can be trained to consistently judge performance on complex tasks. Rather, task-sampling variability is the major source of measurement error. Large numbers of tasks are needed to get a reliable measure of mathematics and science achievement at the elementary level, or to get a reliable measure of job performance in the military. With respect to convergent validity, results suggest that methods do not converge. Performance scores, then, are dependent on both the task and method sampled. (Contains 36 references.) (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED359229

National Center for Research on
Evaluation, Standards, and Student Testing

Final Deliverable – January 1993

Project 2.4: Design Theory and Psychometrics for Complex Performance Assessment in Science

Report on Status of Generalizability Performance: Generalizability and Transfer of Performance Assessments

Sampling Variability of Performance Assessments

Richard J. Shavelson, Project Director

U.S. Department of Education
Office of Educational Research and Improvement
Grant No. R117G10027 CFDA Catalog No. 84.117G

Center for the Study of Evaluation
Graduate School of Education
University of California, Los Angeles
Los Angeles, CA 90024-1522
(310) 206-1532

TM019888

BEST COPY AVAILABLE

The work reported herein was supported in part under the Educational Research and Development Center Program cooperative agreement R117G10027 and CFDA catalog number 84.117G as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the position or policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

SAMPLING VARIABILITY OF PERFORMANCE ASSESSMENTS

Richard J. Shavelson, Xiaohong Gao and Gail P. Baxter*

University of California, Santa Barbara/CRESST

Abstract

In this paper, performance assessments are cast within a sampling framework. A performance assessment score is viewed as a sample of student performance drawn from a complex universe defined by a combination of all possible tasks, occasions, raters, and measurement methods. Using generalizability theory, we present evidence bearing on the generalizability (reliability) and convergent validity of performance assessments sampled from a range of measurement facets, measurement methods, and data bases. Results at both the individual and school level indicate that rater-sampling variability is not an issue: raters (e.g., teachers, job incumbents) can be trained to consistently judge performance on complex tasks. Rather, task-sampling variability is the major source of measurement error. Large numbers of tasks are needed to get a reliable measure of mathematics and science achievement at the elementary level, or to get a reliable measure of job performance in the military. With respect to convergent validity, results suggest that methods do not converge. Performance scores, then, are dependent on both the task and method sampled.

* Now at the University of Michigan.

SAMPLING VARIABILITY OF PERFORMANCE ASSESSMENTS[1,2]

Richard J. Shavelson, Xiaohong Gao and Gail P. Baxter[3]

University of California, Santa Barbara/CRESST

Performance assessments have become political instruments for educational reform in America (Bush, 1991; see also Jaeger, 1992; Shavelson, Baxter, & Pine, 1992; Shavelson, Carey, & Webb, 1990). At state and national levels (e.g. California State Department of Education, 1989; National Assessment of Educational Progress, 1987), a wide range of assessments including open-ended mathematics questions (Pandy, 1991), language-arts writing samples and portfolios (Candell & Ercikan, 1992; Vermont Department of Education, 1991), and hands-on science investigations (Baron, 1990; Camplin, 1989) are being experimented with. The intent is to develop measures of student achievement that focus on students' ability to apply their conceptual understanding and problem-solving skills in novel situations. If assessment systems focus on "higher-order thinking," the reasoning goes, curriculum and teaching can be changed (e.g., Resnick & Resnick, 1992), and ultimately the bottom line—achievement—will be improved.

Heretofore several factors have mediated against the use of performance-based assessments in large-scale testing: cost, time, reliance on complex human judgment of questionable reliability, and lack of convergent or discriminant validity. Judging from current reform, the first two (cost and time) are no longer viewed as barriers, at least for the time being (but see Shavelson, Baxter, & Pine, 1992). States such as California, Connecticut, Maryland, and Vermont are experimenting with performance assessments that are clearly more costly than multiple-choice achievement tests. The second two (reliability and validity) have, to date, received little attention. The purpose of this report is to present empirical evidence on some aspects of the

---

[1] A version of this report focusing just on educational data sets has been accepted for publication by the *Journal of Educational Measurement*.

[2] We wish to thank Dr. Dale Carlson of the California Assessment Program for providing the 1990 Science Performance Assessment data.

[3] Now at the University of Michigan.

technical qualities of performance assessments in elementary mathematics and science.

We view a performance assessment as a concrete, goal-oriented *task* (e.g., discover the contents of a "mystery box" by constructing an electric circuit to it) performed by a student on a particular *occasion* (e.g., sometime in the spring) and evaluated by an expert *rater* who takes into account the process of accomplishing the task as well as the final product. The *method* of presenting the task might be pencil and paper such as an open-ended mathematics problem (e.g., Pandy, 1991), or computer such as a simulation of a science investigation (e.g., Pine, Baxter, & Shavelson, 1991), or laboratory equipment with experts rating performance either in real-time observation or from students' lab notebooks (e.g., Baxter, Shavelson, Goldman, & Pine, 1992).

More specifically, we view a performance assessment score as a sample of student performance drawn from a complex universe defined by a combination of all possible tasks, occasions, raters, and measurement methods. We view the task facet to be representative of the content in a subject-matter domain. The occasion facet includes all possible occasions on which a decision maker would be equally willing to accept a score on the performance assessment. We view the rater facet to include all possible individuals who could be trained to score performance reliably. These three facets traditionally are thought of as sources of unreliability in a measurement.

In addition, we incorporate a method facet into our definition of the universe of generalization. This formulation moves us beyond reliability into a sampling theory of validity (Kane, 1982). Specifically, we view the method facet to be all possible methods (e.g., short-answer, computer simulation) that a decision maker would be equally willing to interpret as bearing on student achievement.

Specification of the task domain is especially critical in measuring achievement in a subject matter: Using performance on a sample of tasks, to what domain does the decision maker generalize? One possible way to link tasks to the broader domain is suggested by Baxter, Shavelson, Herman, Brown, & Valadez (in press; see also Shavelson, Gao, & Baxter, 1992). They linked curricular goals as expressed in the *California State Mathematics Framework* (California State Department of Education, 1985, 1987) with

teaching activities commonly used by teachers in the California Mathematics Project, and translated a sample of these activities into assessments. To be used as an assessment, a goal was set for each activity. For example, ask the student to: (a) find a problem to be solved with the activity, (b) establish criteria by which he/she would know when the problem was successfully solved, or (c) translate among alternative symbolic representations, recognizing their equivalence. This sample of activities was then translated into assessments through an iterative process of development, tryout, modification and tryout.

Student performance may vary across a sample of assessment tasks, raters, occasions, or methods. When performance varies substantially from one *task* sample to another, or from one *occasion* sample to another, or from one *rater* sample to another, we speak of measurement error due to sampling variability. When performance varies from one measurement *method* (e.g., observed performance, computer simulation, short-answer question) to another, we speak of sampling variability due to lack of convergent validity.

Once conceived as a sample of performance from a complex universe, the statistical framework of generalizability (G) theory can be brought to bear on the technical quality of performance-assessment scores (cf. Cronbach, Gleser, Nanda, & Rajaratnam, 1972; see also Brennan, 1991; Kane, 1982; Shavelson, Webb, & Rowley, 1989). From the G theory perspective, an assessment score or profile is but one of many possible samples from a large domain of assessments defined by the particular task, occasion, rater, measurement method (etc.). The theory focuses on the magnitude of sampling variability due to tasks, raters (etc.) and their combinations, providing estimates of the magnitude of measurement error in the form of variance components. Second, it provides a summary coefficient reflecting the "reliability" of generalizing from a sample score or profile to the much larger domain of interest. This coefficient is called a "generalizability" coefficient in G theory, recognizing that generalization may be across different facets, depending on how a performance assessment score is used. The theory also can be used to estimate the magnitude of variability among scores due to method sampling, thereby providing an index of the degree to which alternative measurement methods converge (Kane, 1982).

From a generalizability perspective, sampling variability due to raters, for example, speaks to a traditional concern about the viability of performance

assessments, namely, interrater reliability (cf. Fitzpatrick & Morrison, 1971). Sampling variability due to tasks speaks to the complexity of the subject-matter domain for students. Traditionally, task sampling has been thought of as internal consistency reliability. The goal of test developers has been to make "items" homogeneous to increase reliability. Within the sampling framework, task-sampling variability is dealt with not by homogenizing the tasks but by increasing sample size from the subject-matter domain of interest (cf. Shavelson, Gao, & Baxter, 1992). Sampling variability due to occasions corresponds to the classical notion of retest reliability. From a sampling perspective, it reminds us that decision makers are willing to generalize a student's performance on one particular occasion to many possible occasions. Finally, sampling variability due to measurement method bears on convergent validity (cf. Kane, 1982). Large method sampling variability indicates that measurement methods do not converge, as has commonly been assumed in arguing for the cost-efficiency of multiple-choice testing.

Initially, technical evaluation of performance assessments focused primarily on the impact of rater sampling. With the complexity of performance assessments, the concern was that raters would be inconsistent in their evaluations. More recently, task-sampling variability—inconsistencies in performance across tasks—has been of concern (Dunbar, Koretz, & Hoover, 1991). The findings are remarkably consistent across very diverse studies such as writing, mathematics, and science achievement of elementary students (Baxter et al., in press; Dunbar et al., 1991; Shavelson, Baxter, & Pine, 1991) and job performance of military personnel (Shavelson, Mayberry, Li, & Webb, 1990; Wigdor & Green, 1991a, 1991b). Interrater reliability is not a problem, but task-sampling variability is. Large numbers of tasks are needed to get a generalizable measure.

As our sampling framework suggests, defining the universe of generalization solely in terms of tasks and/or raters is limited. With complex performance measures, a student's achievement score may be impacted by several sources of sampling variability, some associated with generalizability ("reliability")—task, rater, and occasion sampling—and others with convergent validity—method sampling. It therefore becomes important to estimate, simultaneously, as many potential sources of error—task, rater,

occasion, and their interactions—and as many potential sources of method variation—methods and their interactions—as is feasible.

This report, then, presents evidence on the generalizability and convergent validity of performance assessments using data from G studies that sampled a wide range of measurement facets and measurement methods. The data are taken primarily from studies in elementary science and mathematics. Collateral evidence on the generalizability studies for mathematics and science is provided from data on military job performance (Wigdor & Green, 1991a). We draw three education studies: (a) a science assessment study funded by the National Science Foundation (Science); (b) a mathematics assessment study funded by the Office of the President of the University of California (Math); and (c) a science assessment conducted by the California Assessment Program (CAP). These studies were chosen because they provide concrete examples of the impact of various combinations of facets and/or measurement methods on the generalizability and/or convergent validity of students' performance scores (Brennan, 1991; Kane, 1982). The job-performance data were collected in a congressionally mandated study of hands-on job performance of military enlistees. One data set was provided by the Navy (Webb, Shavelson, Kim, & Chen, 1989), the other by the Marine Corps (Shavelson, Mayberry, Li, & Webb, 1990). Findings from all five data sets illustrate the consistency of variance component estimates at the individual (and school) level.

Data Sets and Analyses

Table 1 provides an overview of the data sets used, the questions asked of the data, and the G-study designs. Table 2 presents the formulas for determining generalizability and validity coefficients.

Science

A team of researchers and scientists from the University of California, Santa Barbara and the California Institute of Technology collaborated in developing and evaluating three tasks that were administered by four alternative measurement methods: (a) expert observations of student performance on the hands-on tasks; (b) notebooks, in lieu of observations, in which students recorded their procedures and findings in the hands-on tasks;

Table 1

Overview of Data Sets, Research Questions, and Designs

Data Sets	n_p	Research Questions	Design
Science	26	What is the relative impact of sampling due to raters (r), tasks (t), and occasions (o)?	p x r x t x o
	50	What is the relative impact of sampling due to raters and tasks and how does this compare to findings in math?	p x r x t
	186	What is the relative impact of sampling due to subtasks or "items" within a domain and how does this compare to findings in math?	p x i
		What is the relative impact of sampling due to tasks and methods (m)?	p x t x m
		What is the convergent validity of multiple measurement methods?	p x m
Math	105	What is the relative impact of sampling due to raters and tasks and how does this compare to findings in science?	p x r x t
		What is the relative impact of sampling due to subtasks or "items" within a domain and how does this compare to findings in science?	p x i
CAP	120	What is the relative impact of sampling due to raters and tasks and how does this compare to findings in math?	p x r x t
	120 ($n_{p:s}$=8; n_s=15)	What is the relative impact of sampling due to persons (p), raters (r), and tasks (t) in measuring school-level achievement?	p:s x r x t
Marine Corps	150	What is the relative impact of sampling due to raters and tasks and how does this compare to findings in education?	p x r x t
Navy	26	What is the relative impact of sampling due to raters and tasks and how does this compare to findings in education?	p x r x t

(c) computer simulations of tasks in which students manipulate icons on a Macintosh; and (d) short-answer problems where students answer questions dealing with planning, analyzing, or interpreting the tasks (Shavelson et al., 1991).

Table 2

Equations for Relative and Absolute Generalizability Coefficients

Design	Relative G Coefficient (ρ^2)	Absolute G Coefficient (ϕ)
p x r x t x o*	$\dfrac{\sigma^2_p}{\sigma^2_p + \sigma^2_\delta}$	$\dfrac{\sigma^2_p}{\sigma^2_p + \sigma^2_\Delta}$
p x r x t	$\dfrac{\sigma^2_p}{\sigma^2_p + \dfrac{\sigma^2_{pr}}{n'_r} + \dfrac{\sigma^2_{pt}}{n'_t} + \dfrac{\sigma^2_{prte}}{n'_r n'_t}}$	$\dfrac{\sigma^2_p}{\sigma^2_p + \dfrac{\sigma^2_r}{n'_r} + \dfrac{\sigma^2_t}{n'_t} + \dfrac{\sigma^2_{pr}}{n'_r} + \dfrac{\sigma^2_{pt}}{n'_t} + \dfrac{\sigma^2_{rt}}{n'_r n'_t} + \dfrac{\sigma^2_{prt,e}}{n'_r n'_t}}$
p x t x m	$\dfrac{\sigma^2_p}{\sigma^2_p + \dfrac{\sigma^2_{pt}}{n'_t} + \sigma^2_{pm} + \dfrac{\sigma^2_{ptm,e}}{n'_t}}$	
p:s x r x t**	$\dfrac{\sigma^2_s}{\sigma^2_s + \sigma^2_\delta}$	$\dfrac{\sigma^2_s}{\sigma^2_s + \sigma^2_\Delta}$

Note.

*Relative Error: $\sigma^2_\delta = \dfrac{\sigma^2_{pr}}{n'_r} + \dfrac{\sigma^2_{pt}}{n'_t} + \dfrac{\sigma^2_{po}}{n'_o} + \dfrac{\sigma^2_{prt}}{n'_r n'_t} + \dfrac{\sigma^2_{pro}}{n'_r n'_o} + \dfrac{\sigma^2_{pto}}{n'_t n'_o} + \dfrac{\sigma^2_{prto,e}}{n'_r n'_t n'_o}$

Absolute Error: $\sigma^2_\Delta = \dfrac{\sigma^2_r}{n'_r} + \dfrac{\sigma^2_t}{n'_t} + \dfrac{\sigma^2_o}{n'_o} + \dfrac{\sigma^2_{pr}}{n'_r} + \dfrac{\sigma^2_{pt}}{n'_t} + \dfrac{\sigma^2_{po}}{n'_o} + \dfrac{\sigma^2_{rt}}{n'_r n'_t} + \dfrac{\sigma^2_{ro}}{n'_r n'_o} + \dfrac{\sigma^2_{to}}{n'_t n'_o} + \dfrac{\sigma^2_{prt}}{n'_r n'_t} + \dfrac{\sigma^2_{pro}}{n'_r n'_o} + \dfrac{\sigma^2_{pto}}{n'_t n'_o} + \dfrac{\sigma^2_{rto}}{n'_r n'_t n'_o} + \dfrac{\sigma^2_{prto,e}}{n'_r n'_t n'_o}$

**Relative Error: $\sigma^2_\delta = \dfrac{\sigma^2_{p:s}}{n'_p} + \dfrac{\sigma^2_{sr}}{n'_r} + \dfrac{\sigma^2_{st}}{n'_t} + \dfrac{\sigma^2_{(p:s)r}}{n'_p n'_r} + \dfrac{\sigma^2_{(p:s)t}}{n'_p n'_t} + \dfrac{\sigma^2_{srt}}{n'_r n'_t} + \dfrac{\sigma^2_{(p:s)rt,e}}{n'_p n'_r n'_t}$

Absolute Error: $\sigma^2_\Delta = \dfrac{\sigma^2_r}{n'_r} + \dfrac{\sigma^2_t}{n'_t} + \dfrac{\sigma^2_{p:s}}{n'_p} + \dfrac{\sigma^2_{sr}}{n'_r} + \dfrac{\sigma^2_{st}}{n'_t} + \dfrac{\sigma^2_{rt}}{n'_r n'_t} + \dfrac{\sigma^2_{(p:s)r}}{n'_p n'_r} + \dfrac{\sigma^2_{(p:s)t}}{n'_p n'_t} + \dfrac{\sigma^2_{srt}}{n'_r n'_t} + \dfrac{\sigma^2_{(p:s)rt,e}}{n'_p n'_r n'_t}$

12

One hundred and eighty-six fifth- and sixth-grade students completed each of three science tasks: Paper Towels—conduct an investigation with laboratory equipment to determine which of three paper towels holds, soaks up or absorbs the most/least water; Electric Mysteries—use batteries, bulbs and wires to determine the contents of six black boxes from a list of five possible alternatives (bulb, battery and bulb, wire, two batteries or nothing); Bugs—conduct two experiments to determine sow bugs' preferences for various environments (damp vs. dry, light vs. dark).

Scoring focused on both students' procedures and conclusions, with one exception: Short-answer questions were scored right or wrong. For the Electric Mysteries task, both the circuit used to reach a conclusion as to the contents of a box and the accuracy of the conclusion were taken into account when scoring. For the Paper Towels and Bugs tasks, both the procedures students used to carry out the tasks and their conclusions were taken into account (Baxter et al., 1992). The maximum score for each task was six points.

All tasks were represented by all methods with the exception of the Paper Towels task which could not be adequately simulated. All students were tested on all tasks and all methods. Among those 186 students, a sample of 26 students was administered each observed task and corresponding notebook on two occasions (Ruiz-Primo, Baxter, & Shavelson, in press). In addition, two raters scored a sample of 48 students' Paper Towels and Bugs notebooks.

For the purposes of this paper we draw four examples from this study (see Table 1). The first example (n_p=26), a person x rater x task x occasion G study, examined the relative contribution of raters, tasks, occasions, and their interactions to the generalizability of students' performance scores in elementary science. The second example, a person x rater x task G study (n_p=50), examined sampling variability by comparing results across the Science, Math and CAP data sets (Table 1). The third example examined the relative impact of sampling subtasks or "items" within a domain (e.g., sampling mystery boxes). To this end, p x i G studies were carried out with both science (n_p=186) and math (n_p=105) data. Finally, based on the findings of the second G study that rater variance was negligible, a person (n_p=186) x task x method G study examined the relative contributions made by tasks and methods to sampling variability (cf. Kane, 1982). This study provides evidence

bearing on the convergent validity of measurement methods. Table 2 presents the variance components that enter into the generalizability coefficient for each of these G studies.

Math

Teachers and researchers at the University of California, Santa Barbara developed and evaluated mathematics performance assessments that were closely aligned with hands-on instructional activities (Baxter et al., in press). Hands-on mathematics instructional activities were translated into performance assessments in measurement and place value. In general, the measurement and place-value tasks confronting students were holistic in nature, involved problem solving in concrete situations with the use of manipulatives, and asked students to represent their solutions in various symbolic forms (e.g., written, graphic).

One hundred and five sixth-grade students responded to 7 tasks in the measurement domain and included measurement of length and area. Thirty-one tasks comprised the place-value domain and ranged from arranging numbers according to their place value to a card game in which the largest and smallest sum and difference were constructed, to a translation between chips on a base-ten board and the usual syntactical form of representing place value.

The tasks sampled within each mathematics domain corresponded to "subtasks" or "items" in the science data set. To use comparable terminology for the comparison of findings in mathematics and science, we carried out a person x "item" G study separately for the measurement and the place-value domains (see Table 1).

On three tasks students were asked to respond in writing—2 of 7 tasks in measurement and 1 of 31 tasks in place-value. Within the former, one item asked students to imagine they were talking on the telephone to a friend and to describe a green, 1" x 5" rectangular object such that the friend could draw a picture of it. A second task asked students to justify their choice of a fence perimeter for a dog run to enclose 24 square yards (i.e., 1 x 24, 2 x 12, 3 x 8, 4 x 6). Within the place-value domain, students compared two different representations of the sum of five numbers and explained why they were the same or different (Baxter et al., in press).

All three tasks were scored by two raters using a 6-point holistic scoring rubric (e.g., 1=off track, 4=acceptable, 6=outstanding) developed as part of this study (cf. California State Department of Education, 1989). Here we present data (np=105) from a person x rater x task G study and compare the findings with assessments in science (see Tables 1 and 2).

California Assessment Program (CAP)

The California Assessment Program (CAP) conducted a voluntary, statewide science assessment in 1989-90 with approximately 600 schools. Students were posed five independent tasks. More specifically, students rotated through a series of five self-contained stations at timed intervals (about 15 min). At one station, students were asked to complete a problem-solving task (determine which of these materials may serve as a conductor). At the next station, students were asked to develop a classification system for leaves and then to explain any adjustments necessary to include a new mystery leaf in the system. At yet another, students were asked to conduct tests with rocks and then use the results to determine the identity of an unknown rock. At the fourth station, students were asked to estimate and measure various characteristics of water (e.g., temperature, volume). And at the fifth station, students were asked to conduct a series of tests on samples of lake water to discover why fish are dying (e.g., is the water too acidic?). At each station, students were provided with the necessary materials and asked to respond to a series of questions in a specified format (e.g., fill in a table).

A predetermined scoring "rubric" developed by teams of teachers in California (California State Department of Education, 1990) was used to evaluate students' written responses to each of the tasks. Each rubric was used to score performance on a scale from 0 to 4 (0 = no attempt, 1 = serious flaws, 2 = satisfactory, 3 = competent, 4 = outstanding). All tasks were scored by three raters. For our purposes, we report results of two G studies. The first is a person x rater x task G study carried out for comparison with the Science and Math studies (Tables 1 and 2). The second is a person:school x rater x task design based on data from a random sample of 8 students within each of a random sample of 15 schools scored by 3 raters using the CAP-designed scoring rubric (see Tables 1 and 2).

Collateral Evidence: Military Studies

Marine Corps. One hundred and fifty Marine Corps riflemen were tested by two examiners ("raters") on 35 tasks distributed over 7 stations at two bases (Base A and Base B). The design of the full G study was quite complicated (Shavelson, Mayberry, Li, & Webb, 1990). However, for this report's purpose, the results of a person x rater x task G study are presented here (see Table 1).

Examiners were retired Marine Corps noncommissioned officers (NCOs). The 35 tasks consisted of a variety of job tasks sampled from a large domain that defined the job of rifleman. They included "establishing a helicopter landing zone," "installing a TA-312 telephone set," "performing CPR," "measuring distances on a map," "performing search and safeguard procedures," and "controlling unit firing deployment."

Each task required the performance of from 1 to 36 independent steps. Each step was scored right (1) or wrong (0) by each of two examiners. The total score for a task was calculated as the proportion of steps correctly performed.

Navy. Twenty-six Navy machinist mates were observed by two examiners ("raters") carrying out 11 tasks in a ship's engine room. The examiners were retired noncommissioned officers who had served as machinist mates. The tasks included reading engine gauges, operating equipment, and dealing with casualties. The steps in each task were scored right (1) or wrong (0), and the proportion of steps successfully completed served as the machinist mate's score on that task. A person (n_p=26) x rater x task G study was conducted using this data set (see Table 1).

Results and Discussion

We examined the sampling variability and generalizability of performance assessments at both the individual and school level in a series of G studies. Then we examined method-sampling variability (convergent validity) across several methods of task presentation. Each of the studies presented approximates our conception of a sampling framework that includes raters, tasks, occasions and measurement methods in its definition of performance assessments.

For each study, variance component estimates and generalizability coefficients are presented. Generalizability coefficients are calculated for both

relative decisions—rank ordering students (or schools) or military personnel —and absolute decisions—describing their level of performance. We speak of the former as the "relative G coefficient" and the latter as the "absolute G coefficient." To aid in comparing the magnitude of the variance components, we present the percent of total variability accounted for by each variance component (Shavelson & Webb, 1991).

Generalizability Studies

Sampling variability of performance assessments was examined in a series of G studies. Using the science data, we carried out a person x rater x task x occasion G study to examine the sources of measurement error from a design that comes closest to our notion of a full model (p x r x t x o x m).

To examine the consistency of variance component estimates across subject matter domains, a series of person x rater x task G studies using the Science, Math, and CAP data sets were carried out. Collateral evidence from the military p x r x t G studies is presented as well. To determine the magnitude of variability due to sampling subtasks or "items" within an educational domain (e.g., electricity—variation in performance due to differences in the sample of 6 mystery boxes), a series of p x i(tem) G studies were carried out. Finally, we examined the CAP data in a person:school x rater x task design. We asked: Are variance component estimates and generalizability coefficients similar at the school and individual level?

Individual-level G Studies

G studies were carried out in three different designs approximating the full model. In decreasing order of verisimilitude, they were: (a) person x rater x task x occasion , (b) person x rater x task , and (c) person x item G studies.

Person x rater x task x occasion G study. This G study was carried out with the science data (Table 3). Two raters scored the notebook performance of 26 students who completed two tasks (Bugs and Paper Towels) on two occasions (May and October).

The major source of measurement error was due to the person x task x occasion interaction (59% of the total variability). Some students performed the Paper Towels task better on one occasion but performed the Bugs task more successfully on a different occasion; vice versa for other students. The second

Table 3

Variance Component Estimates for the Person x Rater x Task x Occasion G Study Using the Science Data

Source of Variability	*n*	Estimated Variance Component	Percent Total Variability
Person (p)	26	0.07	4
Rater (r)	2	0.00[a]	0
Task (t)	2	0.00[a]	0
Occasion (o)	2	0.01	1
pr		0.01	1
pt		0.63	32
po		0.00[a]	0
rt		0.00	0
ro		0.00	0
to		0.00[a]	0
prt		0.00[a]	0
pro		0.01	0
pto		1.16	59
rto		0.00[a]	0
prto,e		0.08	4
$(\hat{\rho}^2)$		.04	
$(\hat{\phi})$		.04	

[a]A negative variance component was set to zero.

largest source of error variance was the person x task interaction (32% of the total variability). Students' mean performance scores across raters and occasions depended on the particular task sampled. Some students successfully completed the Paper Towels task but performed less well on the Bugs task; vice versa for other students. Consistent with findings in military job performance (Wigdor & Green, 1991a, 1991b), mathematics performance

(Lane, Stone, Ankenmann, & Lui, 1992), and writing achievement (e.g., Dunbar et al., 1991), large numbers of tasks may be needed to get a generalizable measure of performance.

The magnitude of the other estimated sources of error was negligible and some components were negative. Negative estimates can arise from sampling error when the true value of the component is close to or equal to zero, or from a misspecification of the measurement model (Shavelson et al., 1989). In the present study, negative variance component estimates were due to very little variability among the means of the conditions in each facet. The means averaging across two raters and two occasions were 4.43 and 4.49 for Paper Towels and Bugs, respectively; the means for each rater averaging across two occasions and two tasks were 4.48 and 4.44. Following Brennan (1991), all negative variance component estimates were set to zero.

The generalizability coefficients for relative ($\hat{\rho}^2$) and absolute ($\hat{\phi}$) decisions across raters, occasions, and tasks were relatively low if only one task, rater, and occasion were sampled to form the measurement (.04). Such low generalizability was due to (a) the homogeneous sample of students who, on average, scored quite high on the tasks thereby restricting the range of scores, and (b) large measurement error attributable to the person x task and person x task x occasion interactions.

Person x rater x task G studies. We next examined whether the magnitude of the effect of each measurement facet (raters and tasks) and combinations of these facets was consistent across different assessments and subject domains. To this end, five person x rater x task G studies were carried out (Figure 1). The person x task interaction was consistently the major source of measurement error accounting for 82%, 49%, and 48% of the total variability for the Science, Math and CAP data, respectively.[4] These magnitudes are not unlike those observed with the Navy and Marine Corps data (respectively): 60% and 55% of the variation in job performance was due to the person x task interaction. Once again, task-sampling variability was the major source of error in the performance assessments. For all data sets, the

[4] The difference in the magnitude of the person x task interaction in the Science (82%) and CAP (48%) studies, both involving science assessments, may be due to differences in tasks. CAP tasks prescribed steps in carrying out the tasks; not so the Science study. Or the difference may be due to differences in student populations. Or it may be due to a combination of these factors and some others.

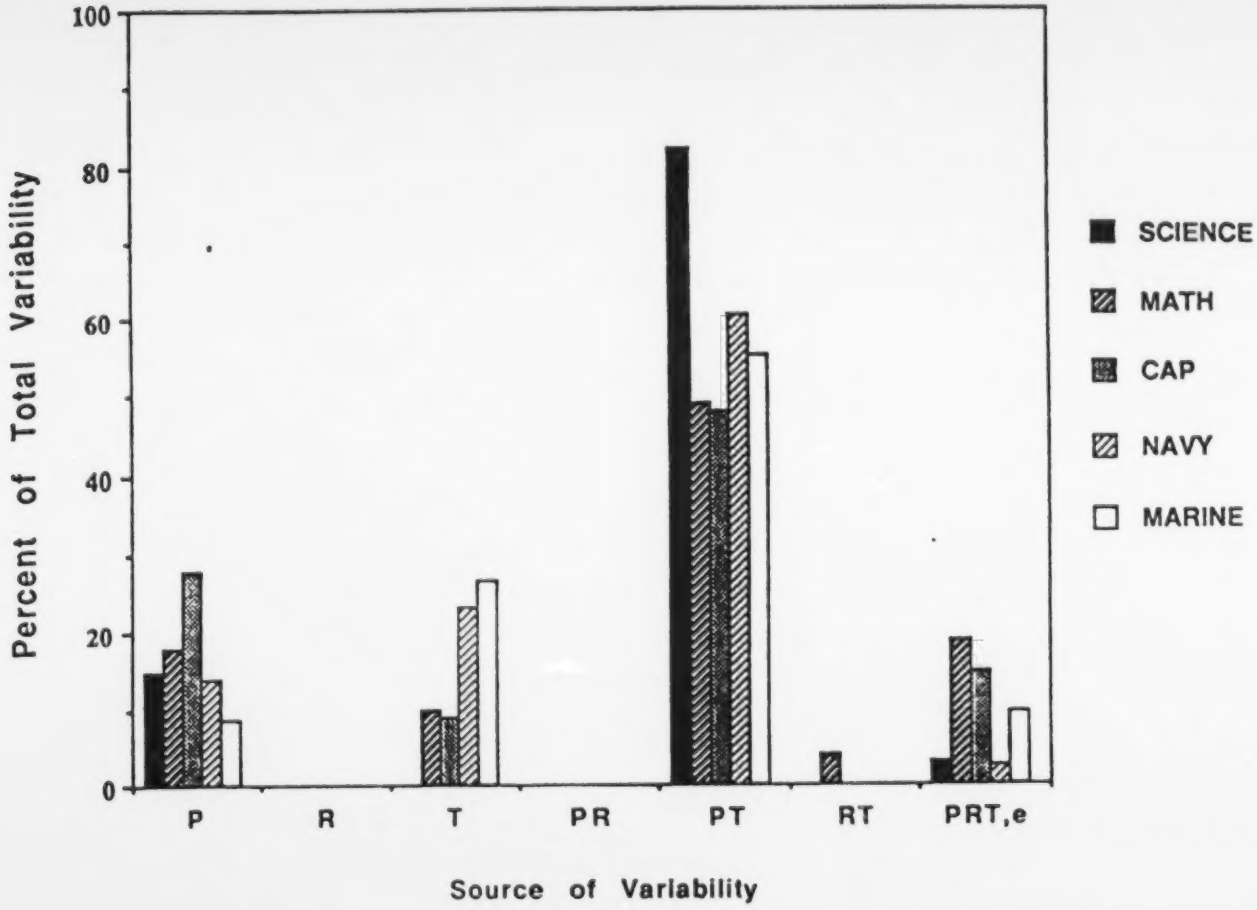


Figure 1. Comparison of the sources of variability in performance assessments across data sets (p x r x t design).

variance components for rater, person x rater, and task x rater interactions were either zero or negligible. In other words, sampling variability due to raters does not appreciably increase measurement error or decrease generalizability. Therefore, one well-trained rater may be sufficient to score performance assessments in mathematics and science, or in the military.

The relative G coefficients were 0.15, 0.21, and 0.32 for the Science, Math, and the CAP studies respectively when only one rater and one task were used. These G coefficients were similar to those reported for the Navy and Marine Corps data, .19 and .12, respectively. The absolute G coefficients were the same or only slightly lower (0.15, 0.18, and 0.29, for education, respectively; .14 and .09 for the military, respectively). To reach an approximate .80 relative G coefficient, about 23 tasks would be needed for the Science study, 15 tasks for the Math study, only 8 tasks for the CAP study, 17 tasks for the Navy study, and 35 tasks for the Marine Corps study.

Person x item G studies. To examine the sampling variability due to items sampled within a domain (e.g., mystery boxes in the electricity domain), we carried out a series of p x i G studies, using Measurement and Place Value data from the math study, and Electric Mysteries and Bugs data from the science study. (Perhaps a rough analogy to this study is to examine the internal consistency of a multiple-choice test in science with a person x item reliability—"internal consistency"—study.) The findings were quite consistent (Figure 2). The major source of error, even within a particular domain (e.g., mystery boxes, several "bugs" experiments), was due to the person x item interaction, once again. (Note that this interaction is confounded with other disturbances not included in the p x i design, and random error.)

School-level. Large-scale assessments may report not only individual student scores but also mean scores for schools and districts. Hence, we would like to know: (a) the main sources of measurement error when schools are the objects of the measurement; (b) the generalizability of an assessment in evaluating school-level achievement; (c) the number of students to be sampled per school and the number of raters and tasks per student needed to get generalizable measures; and (d) the consistency of each facet's effect at the individual and school level.

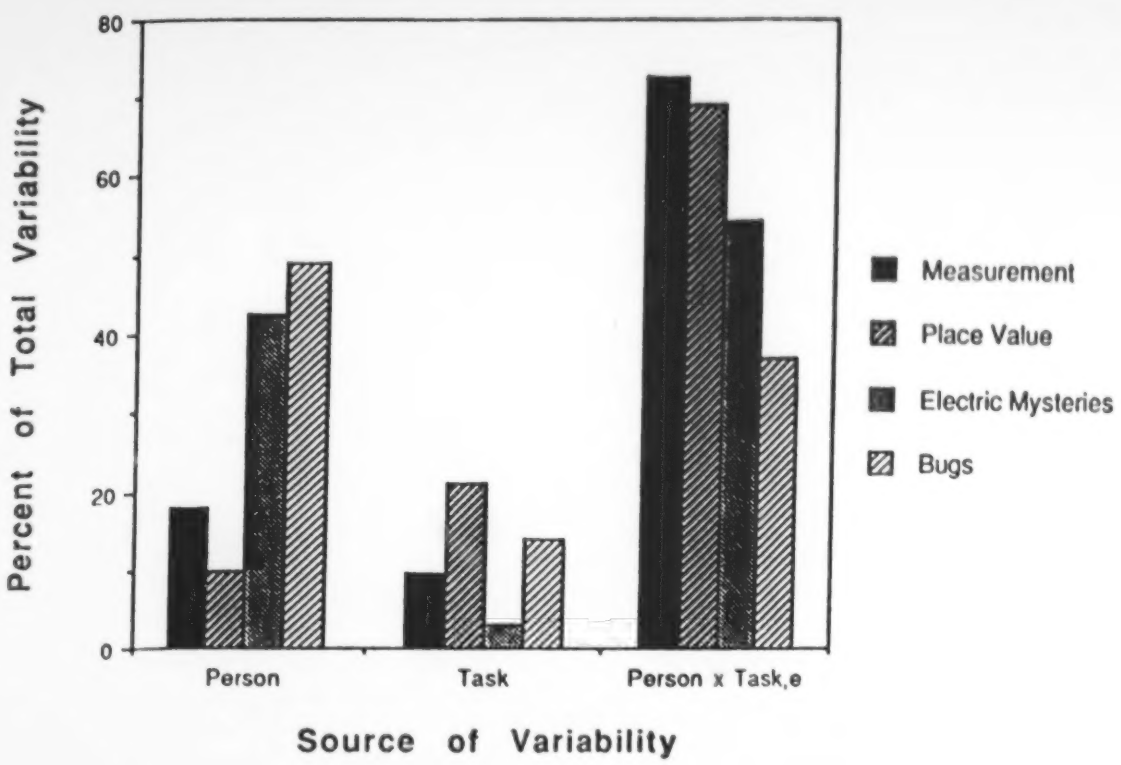


Figure 2. Comparison of the sources of variability in performance assessments across data sets (p x t design).

Following Cardinet, Tourneur, and Allal (1976, 1981) we carried out a person:school x rater x task G study using CAP data (see Table 4). Recall that three raters scored the performance of 8 students within each of 15 schools on 5 science tasks. The major source of measurement error was the person:school x task interaction, accounting for 41% of the total variability in performance. This result, like those reported above at the individual level, indicates that students' performances were inconsistent across the sample of different science tasks; some tasks were easy to do for some students but not for other students.

Variance due to persons (students) is considered measurement error when schools are the objects of measurement. The magnitude of this variance component overshadows the remaining sources of error in the present study, accounting for 22% of the total variability. Variation among students within a

Table 4

Variance Component Estimates for the Person:School x Rater x Task G Study Using the CAP Data

Source of Variability	*n*	Estimated Variance Component	Percent Total Variability
School (s)	15	0.07	7
Rater (r)	3	0.00	0
Task (t)	5	0.09	9
Person:School (p:s)	8	0.23	22
sr		0.00[a]	0
st		0.07	7
rt		0.00	0
(p:s)r		0.00	0
(p:s)t		0.43	41
srt		0.01	1
(p:s)rt,e		0.14	13
$(\hat{\rho}^2)$		0.08	
$(\hat{\phi})$		0.07	

[a]A negative variance component was set to zero.

school was much larger than systematic variation among schools, which accounted for only 7% of the total variability. This finding is consistent with the class-level analysis of Shavelson, Gao, and Baxter (1992).

The estimated variance due to tasks and the task x school interaction accounted for 9% and 7% of the total variability, respectively. Some tasks were more difficult than others across all schools. Furthermore, the average performance scores of some schools on certain tasks were higher than on other tasks.

The variance components for rater, school x rater, task x rater, and person:school x rater rounded to zero. The school x rater x task interaction accounted for about 1% of the total variability. These findings demonstrate that

sampling variability due to raters is not a problem for school-level assessments, but task-sampling variability is.

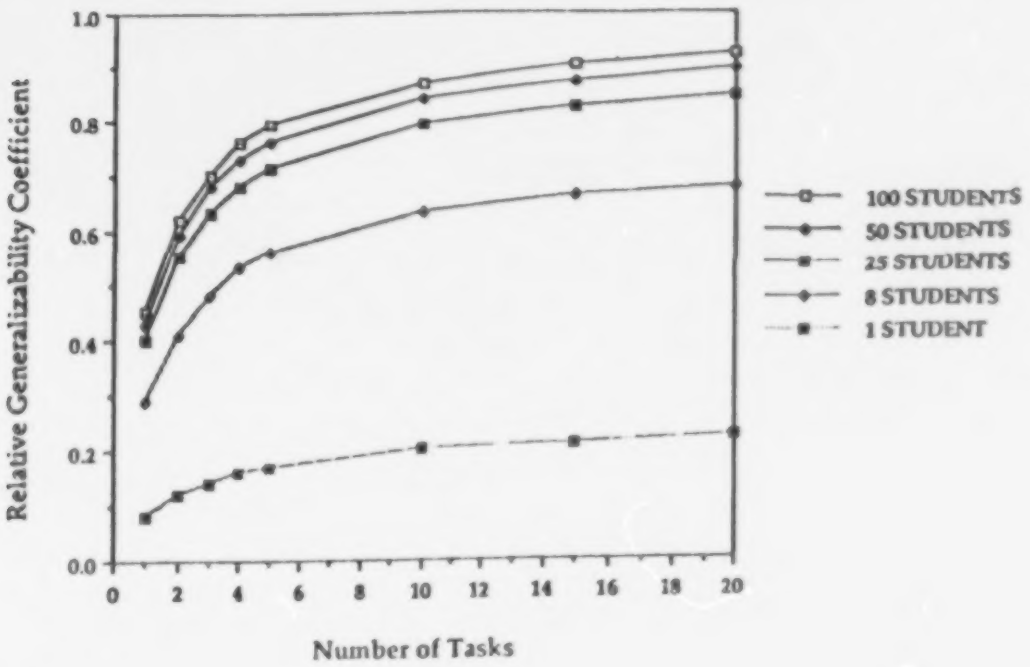
Consequently, we examined the effects of increasing the numbers of tasks on the generalizability of the measure with one rater in a series of decision (D) study designs. Due to the large variability among the students within a school and the large person:school x task interaction, increasing the numbers of students sampled within a school and/or the numbers of tasks produced higher generalizability coefficients for both relative and absolute decisions (Figure 3). To reach generalizability of approximately .80 in estimating a school's mean science achievement regardless of other schools' performances (i.e., absolute decision), a sample of about 50 students within a school would need to be tested on 15 tasks; or about 100 students on 12 tasks. For rank ordering schools, however, only 25 students within a school and 10 tasks or 100 students and 5 tasks would be needed to reach .80 generalizability. In the final analysis, decisions about how many students should participate in the test and how many tasks should be used need to be based on considerations of time, cost, and personnel requirements necessary to develop and administer a test.

Convergent Validity Studies

The validity of performance assessments, specifically the convergent validity of measurement methods, was addressed within the context of Kane's (1982) extension of G theory. The questions raised were: (a) To what extent do the achievement estimates for individual students depend on the particular tasks and/or methods sampled? and (b) Do measurement methods converge in assessing students' science achievement?

One hundred eighty-six students were tested on 2 tasks (Electric Mysteries and Bugs) by each of 4 methods (observed, notebook, computer, and short-answer). A p x t x m G study was carried out and convergent validity was estimated for the average score based on two tasks. This G study, then, examined the convergence of the four measurement methods across two tasks (Electric Mysteries and Bugs).

An examination of the variance component estimates provided in Table 5 for the p x t x m G study indicated that the residual term (p x t x m, e) accounted for the largest portion of the variability in performance scores (29%).



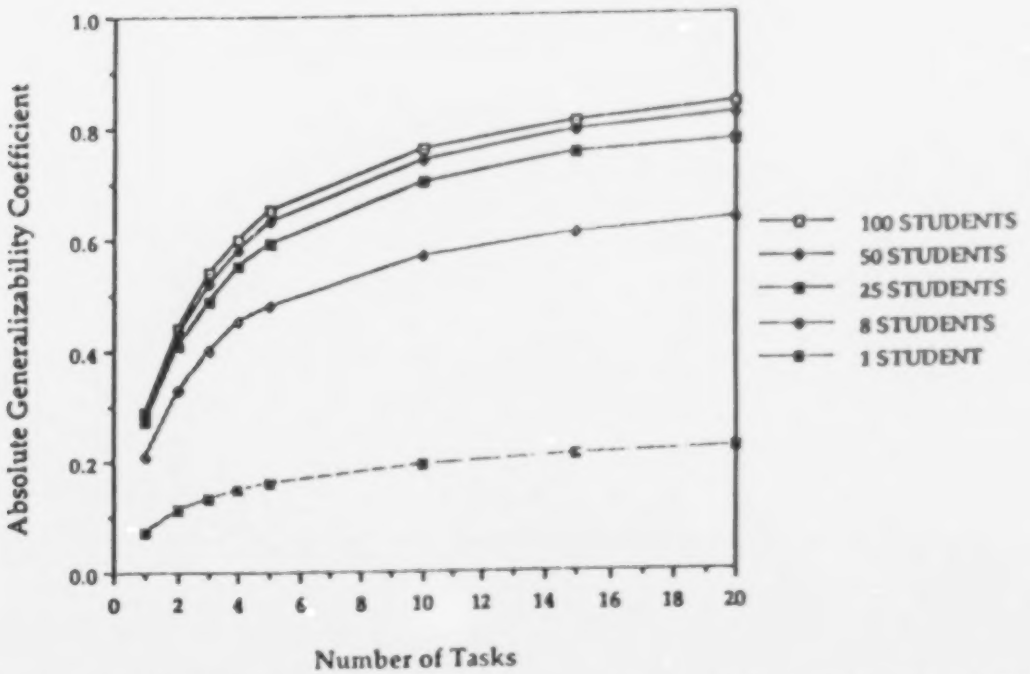


Figure 3. Trade-offs between numbers of tasks and numbers of students needed to achieve various levels of relative (a) and absolute (b) generalizability in CAP data.

Table 5

Exchangeability of Methods (Observed, Notebook, Computer, Short-Answer) and Tasks (Electric Mysteries and Bugs)

Source of Variability	*n*	Estimated Variance Component	Percent Total Variability
Person (p)	186	0.84	19
Task (t)	2	0.70	16
Method (m)	4	0.70	16
pt		0.70	16
pm		0.14	3
tm		0.12	3
ptm,e		1.30	29

Some students were more successful at the Bugs task when using observed scores but more successful at the Electric Mysteries task when using notebooks. However, the magnitude of the effect was confounded with other sources of error (e) not explicitly controlled for in the study.

The variance component for the person x task interaction accounted for 16% of the total variability: A particular student's performance (averaging over all methods) depended on the particular task. Task variability also accounted for 16% of the total variability reflecting the relative difficulty of the tasks; averaging across students and methods, the Bugs task was easier than the Electric Mysteries task.

The method effect accounted for 16% of the total variability in scores. In general, students performed best, on average across tasks, when they used computer simulations (3.90); they had lower scores on the short-answer questions (1.85) than on the other two methods (3.35 for observed and 3.21 for notebooks).

For the particular tasks (Electric Mysteries and Bugs), the average convergent validity coefficient between any pair of randomly sampled methods was .42. This convergent validity coefficient includes a high correlation between direct observation and notebooks—r = .84 for Electric Mysteries and .71

for Bugs—and moderate to low correlations between these two methods and computer simulation and short-answer methods. These findings may be interpreted as indicating that not all methods converge with one another. Rather, the evidence is that certain methods may measure different aspects of achievement (cf. the p x m, e residual).

Conclusions

Development and widespread use of performance-based assessments have not, for the most part, been accompanied by systematic evaluation of their technical qualities. In this paper we bring evidence to bear on the generalizability of performance assessment scores from data sets as diverse as elementary science and mathematics education and military jobs. Moreover, we examine the convergent validity of performance assessments in elementary science. By viewing these assessments within a sampling framework, generalizability theory is used to: (a) estimate potential sources of measurement error or lack of convergence of measurement methods, (b) calculate the generalizability of the measurement, and (c) project alternative designs for collecting large-scale assessment data.

The finding that measurement error is introduced largely by task-sampling variability, and less so by other measurement facets, is consistent with those reported elsewhere for writing achievement and corroborated by our findings for military job performance. Regardless of the subject matter (mathematics or science), domain (education or job performance), or the level of analysis (individual or school), large numbers of tasks are needed to get a generalizable measure of performance.

The finding that a large number of tasks is needed for performance assessments is disquieting to those who would move the testing reform ahead (Rothman, 1992). Increasing the number of tasks is costly and time consuming. Yet this finding should not be so surprising. Multiple-choice science achievement tests typically sample 40 items to get a reliable measure. Nevertheless, assessment reformers attempt to explain the finding away. For example, they claim that the domain of science has been too broadly defined. But this cannot be so if the California Assessment Program constructs an assessment that, it argues, fits well within California's Science Framework and finds large task-sampling variability. Or the claim is made that the tasks

are not sufficiently parallel to one another. But to make tasks parallel to minimize task-sampling variance might make them less representative of the variation among tasks in the domain of interest. Moreover, our research has shown that even parallel tasks (e.g., the six mystery boxes) show considerable person x task sampling variability. In the end, task sampling variability appears to be fact, not artifact. It must be addressed in large-scale assessments.

With regard to convergent validity, results indicate that, at least for the data reported here, student performance is dependent on methods sampled. Methods do not converge. Only notebooks present a reasonable surrogate for observed performance.

Findings of substantial task and method sampling variability have important consequences for the future of performance assessments on a large-scale basis. Generalizations of student or school achievement from a small sample of tasks given by one method to the domain defined by all tasks, raters, occasions, and methods are not supported by the data presented here.

One practical implication of these findings is that, assuming 15 minutes per CAP task, for example, a total of 2.5 hours of testing time would be needed to obtain a generalizable measure (.80) of student achievement. Clearly time and cost are factors to be considered in designing a performance assessment system.

Acknowledgement

This research was funded by grants from the National Science Foundation (SPA-8751511 and TPE-9055443), the University of California President's Office (President's Grant for School Improvement), the U.S. Department of Education through the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) (cooperative agreement R117G10027, CFDA catalog number 84.117G), and the Department of Defense (see Wigdor & Green, 1991a).

Opinions expressed reflect those of the authors and do not necessarily reflect those of the National Science Foundation, the University of California President's Office (President's Grant for School Improvement), or the Department of Defense.

The findings and opinions expressed in this report do not reflect the position or policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

References

Baron, J. B. (1990, October). *Use of alternative assessments in state assessment: The Connecticut experience.* Paper presented at conference on the Promise and Peril of Alternative Assessment, Washington, DC.

Baxter, G. P., Shavelson, R. J., Goldman, S. R., & Pine, J. (1992). Evaluation of procedure-based scoring for hands-on science assessment. *Journal of Educational Measurement, 29*(1), 1-17.

Baxter, G. P., Shavelson, R. J., Herman, S. J., Brown, K. A., & Valadez, J. (in press). Mathematics performance assessment: Technical quality and diverse student impact. *Journal of Research in Mathematics Education.*

Brennan, R. L. (1991). *Elements of generalizability theory* (2nd ed). Iowa City, IA: The American College Testing Program.

Bush, G. W. (1991). *America 2000: An education strategy.* Washington, DC: U.S. Department of Education.

California State Department of Education. (1985). *Mathematics framework for California public schools: Kindergarten through grade twelve.* Sacramento, CA: California State Department of Education, Curriculum Framework and Textbook Development Unit.

California State Department of Education. (1987). *Mathematics framework for California public schools: Kindergarten through grade twelve.* Sacramento, CA: California State Department of Education, Curriculum Framework and Textbook Development Unit.

California State Department of Education. (1989). *A question of thinking: A first look at students' performance on open-ended questions in mathematics.* Sacramento, CA: Author.

California State Department of Education. (1990). *Sixth-grade science performance-based assessment administration manual.* Sacramento, CA: Author.

Camplin, J. (1989, October). *New York state science assessment.* Paper presented at the Curriculum/Assessment Alignment Conference, Long Beach, CA.

Candell, G. L., & Ercikan, K. (1992). *Assessing the reliability of the Maryland school performance assessment program using generalizability theory.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Cardinet, J., Tourneur, Y., & Allal, L. (1976). The symmetry of generalizability theory. Applications to educational measurement. *Journal of Educational Measurement, 13*(2), 119-135.

Cardinet, J., Tourneur, Y., & Allal, L. (1981). Extensions of generalizability theory and its applications in educational measurement. *Journal of Educational Measurement, 18*(4), 183-204.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, H. (1972). *The dependability of behavioral measurements.* New York: Wiley.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*(4), 289-303.

Fitzpatrick, R., & Morrison, E. J. (1971). Performance and product evaluation. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 237-270). Washington, DC: American Council on Education.

Jaeger, R. M. (1992). "World class" standards, choice, and privatization: Weak measurement serving presumptive policy. *Phi Delta Kappan, 74*(2), 118-128.

Kane, M. T. (1982). A sampling model of validity. *Applied Psychological Measurement, 6*, 125-160.

Lane, S., Stone, C. A., Ankenmann, R. D., & Lui, M. (1992, April). *Empirical evidence for the reliability and validity of performance assessments.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

National Assessment of Educational Progress. (1987). *Learning by doing* (Report No. 17-HOS-80). Princeton, NJ: Educational Testing Service.

Pandy, T. (1991). *A sampler of mathematics assessment.* Sacramento: California State Department of Education.

Pine, J., Baxter, G. P., & Shavelson, R. J. (1991, April). *Computer simulations for assessment.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement, and instruction* (pp. 37-75). Boston: Kluwer Academic Publishers.

Rothman, R. (1992). Performance-based assessment gains prominent place on research docket. *Education Week, 12*(9), 1, 22, 24.

Ruiz-Primo, M. A., Baxter, G. P., & Shavelson, R. J. (in press). On the stability of performance assessments. *Journal of Educational Measurement.*

Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessments in science. *Applied Measurement in Education, 4*(4), 347-362.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher, 21*(4), 22-27.

Shavelson, R. J., Carey, N. B., & Webb, N. M. (1990). Indicators of science achievement: Options for a powerful policy instrument. *Phi Delta Kappan, 71*(9), 692-697.

Shavelson, R. J., Gao, X., & Baxter, G. P. (1992). *Content validity of performance assessments: Centrality of domain specification.* Submitted to *Journal of Educational Measurement.*

Shavelson, R. J., Mayberry, P. W., Li, W., & Webb, N. M. (1990). Generalizability of job performance measurements: Marine corps rifleman. *Military Psychology, 2*(3), 129-144.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage Publications.

Shavelson, R. J., Webb, N. M., & Rowley, G. (1989). Generalizability theory. *American Psychologist, 44*, 922-932.

Vermont Department of Education. (1991). *This is my best. Vermont's writing assessment program.* Montpelier, VT: Author.

Webb, N. M., Shavelson, R. J., Kim, K-S., & Chen, Z. (1989). Reliability (generalizability) of job performance measurements: Navy Machinist Mates. *Military Psychology, 1*, 91-110.

Wigdor, A. K., & Green, B. F. (Eds.). (1991a). *Performance assessment in the workplace* (Vol. 1). Washington, DC: National Academy Press.

Wigdor, A. K., & Green, B. F. (Eds.). (1991b). *Performance assessment in the workplace: Technical issues* (Vol. 2). Washington, DC: National Academy Press.